Electronic Journal of Statistics
Vol. 8 (2014) 1714–1723
ISSN: 1935-7524
DOI: 10.1214/14-EJS900A

Analysis of proteomics data: Block k-mean alignment*

Mara Bernardi, Laura M. Sangalli†, Piercesare Secchi and Simone Vantini

MOX – Department of Mathematics, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133, Milano, Italy
e-mail: marasabina.bernardi@mail.polimi.it; laura.sangalli@polimi.it
piercesare.secchi@polimi.it; simone.vantini@polimi.it

Abstract: We analyze the proteomics data introducing a block k-mean alignment procedure. This technique is able to jointly align and cluster the data, accounting appropriately for the block structure of these data, that includes measurement repetitions for each patient. An analysis of area-under-peaks, following the alignment, separates patients who respond and those who do not respond to treatment.

Keywords and phrases: Block k-mean alignment, registration, functional clustering, proteomics data.

Received August 2013.

1. Block k-mean alignment

Motivated by the analysis of the proteomics dataset described in Koch, Hoffman and Marron (2014), we introduce here a variant of the k-mean alignment procedure that accounts appropriately for the block structure of these data. Likewise the k-mean alignment technique described in Sangalli et al. (2010a) and Sangalli, Secchi and Vantini (2014), the proposed block variant is able to jointly align and cluster in k clusters a set of functional data; moreover, it complies with partially exchangeable structures in the data. In the proteomics application, partial exchangeability is due to the presence of measurement repetitions for each subject.

Consider a set of functions composed by repetitions of the same measurement on different subjects or experimental units:

$$\{ f_{ij}(t) \mid i = 1, \dots, m\,; j = 1, \dots, n_i \}\,,$$

where m is the number of experimental units, n_i is the number of exchangeable measurements for the i-th experimental unit, and $f_{ij}(t)$ is the j-th measurement for the i-th experimental unit at time t. The total number of functions is $n = n_1 + \cdots + n_m$. The set of exchangeable measurements for the same experimental

*Main article 10.1214/14-EJS900.
†Corresponding author.

unit, $\{f_{ij}(t)|j = 1,\dots,n_i\}$ for $i = 1,\dots,m$, is referred to as block. In the proteomics dataset the experimental units are the patients.

The block k-mean alignment consists of two concatenated steps: the *within block alignment* and the *between block alignment and clustering*.

1) *Within block alignment*. In this step, each block $\{f_{ij}(t)|j = 1,\dots,n_i\}$, for $i = 1,\dots,m$, is considered independently from the others. The curves within the same block are aligned. To this end, the k-mean alignment algorithm described in Sangalli et al. (2010a) and Sangalli, Secchi and Vantini (2014) is used, with $k = 1$ (see also Sangalli et al., 2009). In fact, since the curves within the same block are replicated measurements, it does not make sense here to consider multiple clusters.

Let $\tilde{f}_{ij}$ be the within block aligned curves.

2) *Between block alignment and clustering*. In this step, the measurements on the same experimental unit are treated in block. The k-mean alignment algorithm is applied to the m blocks of curves

$$\left\{ \left\{ \tilde{f}_{1j}(t) \,\middle|\, j = 1,\dots,n_1 \right\},\dots,\left\{ \tilde{f}_{mj}(t) \,\middle|\, j = 1,\dots,n_m \right\} \right\},$$

so that the curves in the same block are assigned to the same cluster, each curve in the same block being warped with the same warping function.

The total alignment is the composition of the two warping functions found in the two steps, the *within block alignment* and the *between block alignment and clustering*.

Block k-mean alignment allows to explore possible clustering structures among the experimental units. Thanks to its block structure, it avoids incoherent results where the measurement repetitions of the same experimental units are assigned to different clusters.

The analysis here presented have been performed using `fdakma` R package downloadable from CRAN (see Parodi et al., 2014).

2. Block *k*-mean alignment of the proteomics data

In the Proteomics dataset the fifteen curves are actually five blocks of three curves each: a block represents a patient, while the three curves within each block are TIC profile measurement repetitions. In this case $m = 5$, $n_i = 3$ for $i = 1,\dots,5$, and the blocks are given by

$$\{ f_{1j} = A_j \mid j = 1,2,3 \}, \{ f_{2j} = B_j \mid j = 1,2,3 \}, \{ f_{3j} = C_j \mid j = 1,2,3 \},$$

$$\{ f_{4j} = X_j \mid j = 1,2,3 \}, \{ f_{5j} = Y_j \mid j = 1,2,3 \}.$$

A simple scheme of the block k-mean alignment in the case of the Proteomics data is represented in Figure 1.

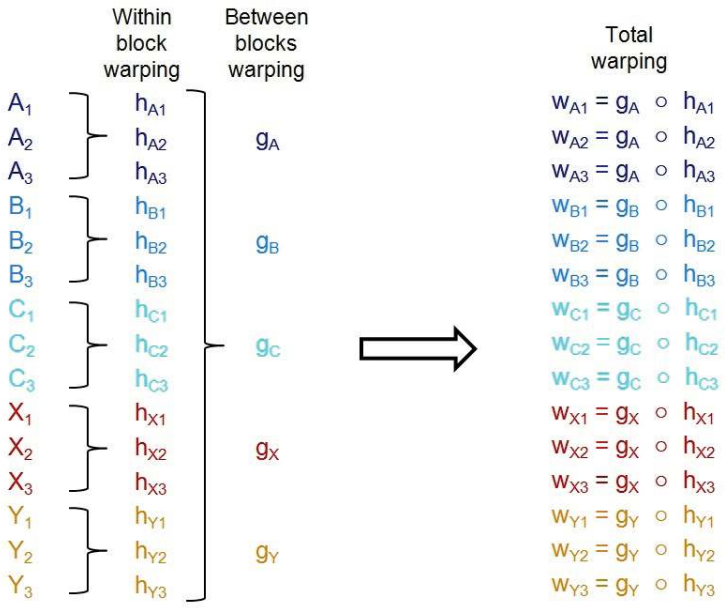

FIG 1. *Scheme of the block k-mean alignment in the case of the Proteomics data.*

We shall consider two TIC profiles to be perfectly aligned if they are identical up to a multiplicative factor. This choice is due to the characteristics of the data, which have the same baseline and differ only for the peak pattern. Indeed, the signature of a TIC profile is given by the relative heights of the peptide peaks. Therefore, we shall use the following similarity index:

$$\rho(f_i, f_j) = \frac{\int f_i(s) f_j(s) ds}{\sqrt{\int f_i(s)^2 ds} \sqrt{\int f_j(s)^2 ds}}. \tag{2.1}$$

Indeed this similarity index assigns maximal similarity (similarity equal to 1) to curves that differ only by a positive multiplicative factor:

$$\rho(f_i, f_j) = 1 \;\Leftrightarrow\; \exists a \in \mathbb{R}^+ : f_i(t) = a f_j(t).$$

The integrals in (2.1) are computed over the intersection of the domains of the curves f_i and f_j.

The physical phenomenon does not suggest a unique group of warping functions to use, hence the analysis were done with different groups of warping functions in order to choose the group that provides the best results on the data. We choose four groups that are coherent with the similarity index chosen.

$$\mathcal{H}_{\text{affine}} = \{h : h(t) = mt + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\},$$
$$\mathcal{H}_{\text{shift}} = \{h : h(t) = t + q \text{ with } q \in \mathbb{R}\},$$

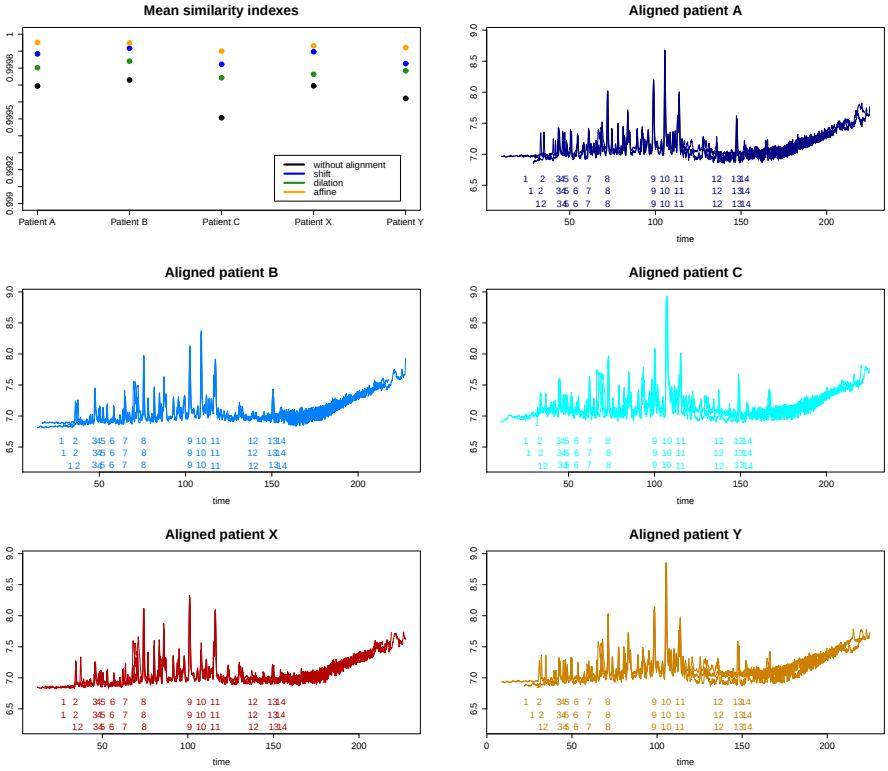

FIG 2. *Results of the* within block alignment. *The first panel shows the mean similarity indexes of the unaligned curves (black) and those obtained after the alignment with* $\mathcal{H}_{shift}$ *(blue),* $\mathcal{H}_{dilation}$ *(green) and* $\mathcal{H}_{affine}$ *(orange). The other panels show the alignment within each block with* $\mathcal{H}_{affine}$ *warping functions.*

$$\mathcal{H}_{\text{dilation}} = \{h : h(t) = mt \text{ with } m \in \mathbb{R}^+\},$$
$$\mathcal{H}_{\text{identity}} = \{h : h(t) = t\},$$

where the last one corresponds to the case where no alignment is indeed performed.

In this analysis the k cluster templates are computed as medoids, i.e., the curves in the sample that maximize the total similarity; see eq. (1.1) in Sangalli, Secchi and Vantini (2014). See Sangalli et al. (2010b) for details. Medoids are in fact more representative of these data that are characterized by numerous sharp peaks.

The first panel of Figure 2 shows the results obtained in the first step, the *within block alignment*. For each of the five patients, the plot shows the means of the similarity indexes between the within block aligned functions and the corresponding within block templates. The black dots represent the means of the similarities between the unaligned data and their within block templates. The

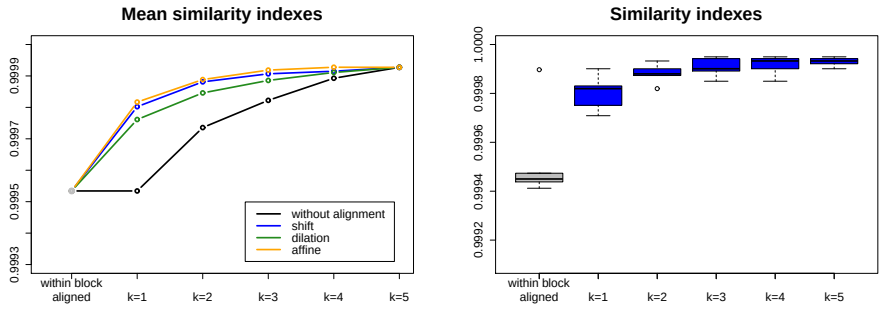

FIG 3. *Results of the* between block alignment and clustering. *The left panel shows the mean similarity indexes between curves and their corresponding templates, considering different number of clusters k and different classes of warping functions. The right panel displays the boxplots of the similarity indexes obtained using the group of warping functions* $\mathcal{H}_{shift}$.

blue, green and orange dots indicate the similarities obtained after the within block alignment respectively with only shift, only dilation and affine warping. The figure shows that for all five patients the highest similarity is obtained using the group of affine warping functions. Hence, in the within block alignment step, we choose the group of warping functions $\mathcal{H}_{\text{affine}}$. The other panels of Figure 2 show the registration thus obtained by within block alignment. The figure shows a good registration within each block, with the three TIC profiles of each patient well aligned. The bottom of each plot displays the retention times of the reference peptides provided with the data. It should be noticed that the retention times of the reference peptides have not been used for the alignment; they are displayed only to show the good alignment results.

The alignment between patients is obtained with the second step: the *between block alignment and clustering*. The results of this step are shown in Figure 3. The left panel shows the means of the similarity indexes between the functions, aligned and clustered between blocks, and their corresponding templates, for different number of clusters k. The gray dot indicates here the mean similarity of the within block aligned curves and their corresponding within block templates. In black the results obtained with the k-mean alignment with no warping allowed ($\mathcal{H}_{\text{identity}}$), i.e., the functional k-mean clustering. In color the results obtained with different classes of warping functions: only shifts in blue, only dilations in green and affine transformations in orange. For $k = 5$ the similarities coincide. Indeed, in this case each cluster coincides with a block of (within block) aligned curves, so that there is no need to further align.

The left panel of Figure 3 shows that the alignment of the functions increases their similarity. The results obtained with the three groups of warping functions, $\mathcal{H}_{\text{shift}}$, $\mathcal{H}_{\text{dilation}}$ and $\mathcal{H}_{\text{affine}}$, are very similar. With only dilation the similarities obtained are slightly lower than those with only shift or affine warping. The similarities obtained with $\mathcal{H}_{\text{shift}}$ and $\mathcal{H}_{\text{affine}}$ are almost identical. We therefore choose to use the group $\mathcal{H}_{\text{shift}}$. The right panel of Figure 3 shows in blue the

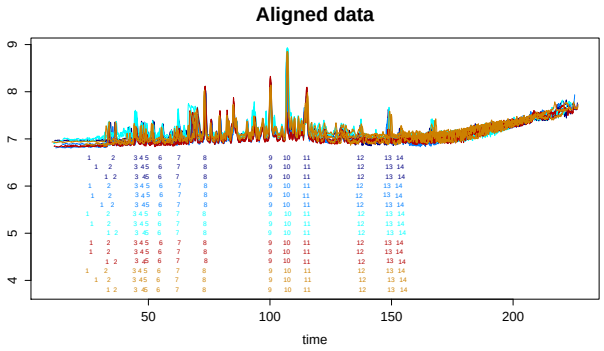

FIG 4. *Total alignment obtained by block k-mean alignment, considering* $k = 1$ *cluster (simple alignment without clustering).*

boxplots of the similarities between the functions, aligned and clustered between blocks with $\mathcal{H}_{\text{shift}}$, and their corresponding templates, for different number of clusters k. The gray boxplot refers to the similarities between the within block aligned functions and their corresponding templates. The variability shown by the boxplot of the case $k = 5$ is the residual variability amongst the within block aligned curves. The total alignment is the composition of the *within block alignment* with the group of warping functions $\mathcal{H}_{\text{affine}}$ and the *between block alignment and clustering* with the group of warping functions $\mathcal{H}_{\text{shift}}$. In the following we describe the results obtained by *between block alignment and clustering* with $k = 1$ and $k = 2$ clusters.

Figure 4 shows the total alignment obtained with $k = 1$ cluster (i.e., simple alignment without clustering). A visual inspection of the aligned data and of the retention times of the reference peptides highlights the very good alignment results. Only the first two reference peptides appear not well aligned. Note that the first reference peptide is not well aligned also by the procedures considered for instance in Cheng et al. (2014), Tucker, Wu and Srivastava (2014) and Lu, Koch and Marron (2014). This peptide is not associated to a peak of the TIC profile and we wonder if its reference identification may have been inaccurate. Also the second peptide proves to be difficult to align even when using the more flexible warping functions considered by Tucker, Wu and Srivastava (2014). As suggested by a Referee, the not so good alignment of the first two reference peptides may also be due to a drift effect caused of the measurement instrument (see Koch, Hoffman and Marron (2014)), as the third measurement from each patient appears to have a truncated spectrum, with the first two identified reference peptides offset to the right. See also Figure 2 that better illustrates this aspect in within block aligned TIC profiles. Figure 5 shows the corresponding total warping functions, colored according to two different criteria. In the left panel the colors refer to the patients (blocks): the three warping functions of the TIC profiles for the same patient have the same color. Instead, the right panel displays the same warping functions colored according to the order of

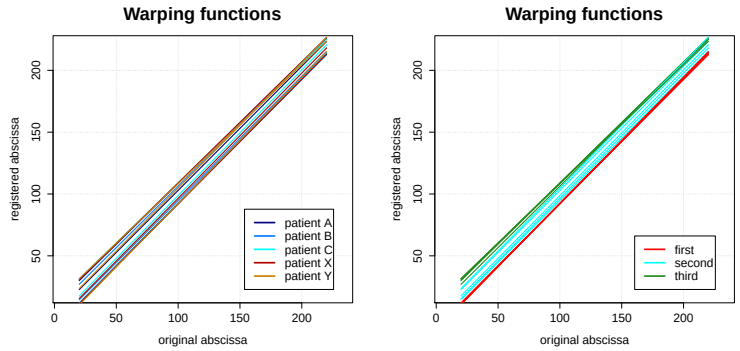

FIG 5. *Total warping functions for the case $k = 1$ colored according to the patient (left panel) and to the order of the TIC profile within the patient (right panel).*

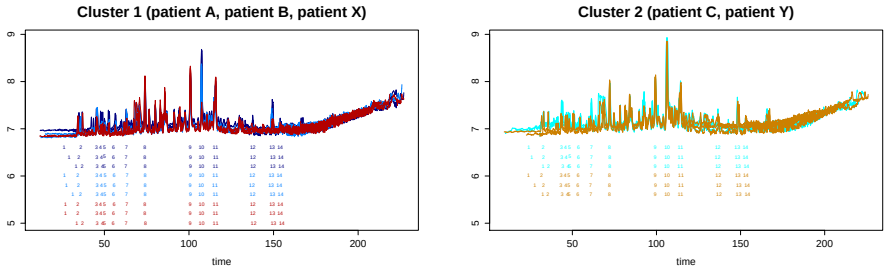

FIG 6. *Total alignment obtained by block k-mean alignment, considering $k = 2$ clusters. The two clusters are represented in two different panels.*

each TIC profile within each patient (block): in red the 5 first TIC profiles for the 5 patients, in light blue the second TIC profiles and in green the thirds. The left panel does not show any clustering of the patients in the phase. This means that phase variability is not related to the patient. Instead, the right panel displays a clear clustering of the warping functions of the first, second and third TIC profiles. The main difference amongst the three groups is the value of the intercept of the warping functions. This phase variability is due to the measuring instrument which introduced a time drift in the TIC profiles, as described in Koch, Hoffman and Marron (2014) and already mentioned above. In order to make up for the measurement drift, all the first TIC profiles must be anticipated, while all the third TIC profiles must be delayed.

We now describe the results obtained considering $k = 2$ clusters in the *between block alignment and clustering* step, hence exploring possible clustering in the amplitude of the TIC profiles. The case $k = 2$ is particularly interesting since a visual inspection of the similarities obtained by setting $\mathcal{H}_{shift}$ as the group of warping functions, displayed in blue in the left and right panels of Figure 3, suggests the existence of $k = 2$ clusters. Figure 6 shows the TIC profiles aligned

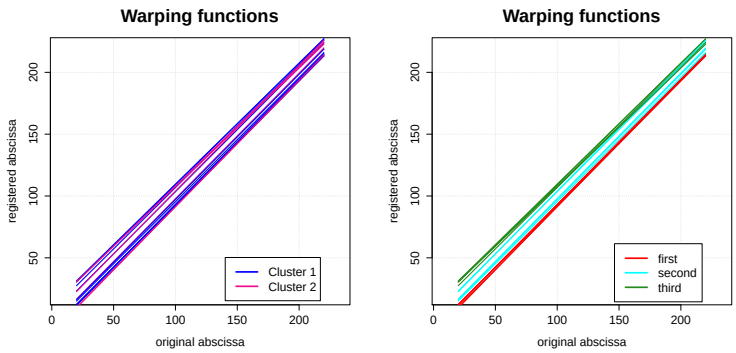

FIG 7. *Total warping functions for the case $k = 2$ colored according to the cluster (left panel) and to the order of the TIC profile within the patient (right panel).*

and clustered in $k = 2$ clusters, displayed in the two panels. The alignment of the curves within both clusters is very good, with only the first and second peptides being problematic, as commented earlier. The first cluster, left panel, is composed of patients A, B and X, while the second cluster, right panel, is composed of patients C and Y.

The left panel of Figure 7 shows the total warping functions, colored according to two clusters. No further clustering is apparent in the phase. Instead, the right panel of the same figure displays the same warping functions colored according to the order of each TIC profile within each patient (block), likewise in the right panel of Figure 5. The same observations made previously, according to clustering in the phase of first, second and third TIC profiles for each patient, still hold.

The clustering in amplitude suggested by the procedure (patients A, B, and X in one cluster and patients C and Y in the other) is not related to response to chemotherapy. This clustering is related to some other feature distinguishing the patients and it would be worthy of further investigation; more needs to be know about the patients for this exploration. We note that performing the analysis without considering the partial exchangeable structure of the data, and applying the k-mean algorithm directly to the fifteen TIC profiles, leads to a very similar clustering result, but with the inconsistency that the third TIC profile of patient A is clustered together with the TIC profiles of patients C and Y. With the block k-mean alignment this inconsistency is avoided.

It is however possible to discriminate patients who respond and patients who do not respond to chemotherapy using for instance area-under-peaks. Suppose that, after the alignment of the TIC profiles, it is possible to identify the reference peptides, for example by comparison to a given template whose reference peptides' retention times are known. We consider the last twelve of the fourteen reference peptides (from the 3^{rd} to the 14^{th}) and exclude instead the first two reference peptides, since they are not well aligned by our procedure. We compute the area under the twelve peaks by fixing a width for each of the twelve

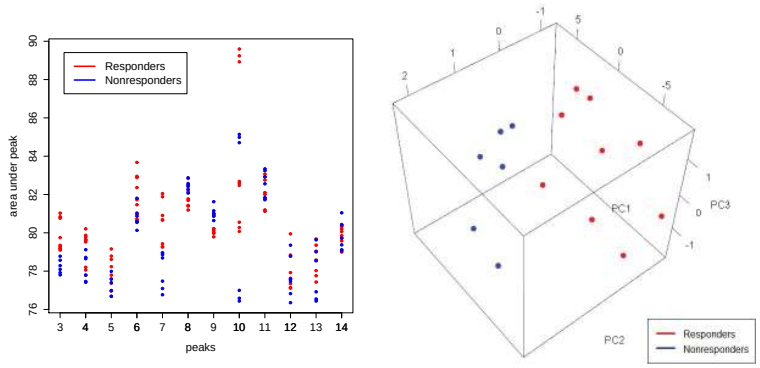

FIG 8. *Discrimination between responders and non-responders using the area under the peaks.*

considered peaks and using that same width for all the fifteen TIC profiles. The left panel of Figure 8 shows the values of the area-under-peaks for the fifteen TIC profiles. The red dots correspond to the patients responding to chemotherapy, the blue ones to the patients who are not responding. Some peaks seem to discriminate well the two groups of patients (for example peak 3 and peak 7). We then performed PCA on area-under-peaks to reduce data dimensionality. The right panel of Figure 8 shows the projections of the data along the first three principal components: responders and not-responders are very well separated. We also run the same analysis on the area-under-peaks after subtracting the baseline to the data, obtaining the same results.

3. Discussion

As highlighted by the analyses, in this application the time warping seems truly affine, and more specifically mainly captured by simple shifts, with phase variability amongst data mostly due to time drifts of the measuring instrument. This and the finding on clustering in the phase of first, second and third TIC profiles, are fully consistent with the data description given in Koch, Hoffman and Marron (2014). In fact, also when using classes of warping functions richer than the group of affinities, improvements in the alignment results are noticed when forcing the warping toward linear (Lu, Koch and Marron (2014)) or toward simple shifts (Cheng et al. (2014)).

Acknowledgements

We are grateful to MBI Mathematical Biosciences Institute http://mbi.osu.edu/ for support. L. M. Sangalli acknowledges funding by the research program Dote Ricercatore Politecnico di Milano – Regione Lombardia, project: Functional data analysis for life sciences, and by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, *FIRB Futuro in Ricerca* starting grant SNAPLE: Statistical and

Numerical methods for the Analysis of Problems in Life sciences and Engineering http://mox.polimi.it/users/sangalli/firbSNAPLE.html.

References

Cheng, W., Dryden, I. L., Hitchcock, D. B. and Le, H. (2014). Analysis of proteomics data: Bayesian alignment of functions. *Electronic Journal of Statistics* **8** 1734–1741, Special Section on Statistics of Time Warpings and Phase Variations.

Koch, I., Hoffman, P. and Marron, J. S. (2014). Proteomics profiles from mass spectrometry. *Electronic Journal of Statistics* **8** 1703–1713, Special Section on Statistics of Time Warpings and Phase Variations.

Lu, X., Koch, I. and Marron, J. S. (2014). Analysis of proteomics data: Impact of alignment on classification. *Electronic Journal of Statistics* **8** 1742–1747, Special Section on Statistics of Time Warpings and Phase Variations.

Parodi, A., Patriarca, M., Sangalli, L. M., Secchi, P., Vantini, S. and Vitelli, V. (2014). fdakma: Clustering and alignment of a given set of curves, R package version 1.1.1.

Sangalli, L. M., Secchi, P. and Vantini, S. (2014). Analysis of AneuRisk65 data: k-mean alignment. *Electronic Journal of Statistics* **8** 1891–1904, Special Section on Statistics of Time Warpings and Phase Variations.

Sangalli, L. M., Secchi, P., Vantini, S. and Veneziani, A. (2009). A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *J. Amer. Statist. Assoc.* **104** 37–48. MR2663032

Sangalli, L. M., Secchi, P., Vantini, S. and Vitelli, V. (2010a). K-mean alignment for curve clustering. *Computational Statistics and Data Analysis* **54** 1219–1233. MR2600827

Sangalli, L. M., Secchi, P., Vantini, S. and Vitelli, V. (2010b). Functional clustering and alignment methods with applications. *Communications in Applied and Industrial Mathematics* **1** 205–224. MR2812260

Tucker, J. D., Wu, W. and Srivastava, A. (2014). Analysis of proteomics data: Phase amplitude separation using extended Fisher-Rao metric. *Electronic Journal of Statistics* **8** 1724–1733, Special Section on Statistics of Time Warpings and Phase Variations.